



# Extreme mitochondrial reduction in a novel group of free-living metamonads

Received: 11 August 2023

Accepted: 15 July 2024

Published online: 09 August 2024

Shelby K. Williams <sup>1,2</sup>, Jon Jerlström Hultqvist <sup>1,3</sup>, Yana Eglit <sup>1,4</sup>, Dayana E. Salas-Leiva <sup>1,5</sup>, Bruce Curtis <sup>1,2</sup>, Russell J. S. Orr <sup>6</sup>, Courtney W. Stairs <sup>7</sup>, Tuğba N. Atalay <sup>1,2</sup>, Naomi MacMillan <sup>1,2</sup>, Alastair G. B. Simpson <sup>1,4</sup> & Andrew J. Roger <sup>1,2</sup>

Check for updates

Metamonads are a diverse group of heterotrophic microbial eukaryotes adapted to living in hypoxic environments. All metamonads but one harbour metabolically altered ‘mitochondrion-related organelles’ (MROs) with reduced functions, however the degree of reduction varies. Here, we generate high-quality draft genomes, transcriptomes, and predicted proteomes for five recently discovered free-living metamonads. Phylogenomic analyses placed these organisms in a group we name the ‘BaSK’ (Barthelonids+Skoliomonads) clade, a deeply branching sister group to the Fornicata, a phylum that includes parasitic and free-living flagellates. Bioinformatic analyses of gene models shows that these organisms are predicted to have extremely reduced MRO proteomes in comparison to other free-living metamonads. Loss of the mitochondrial iron-sulfur cluster assembly system in some organisms in this group appears to be linked to the acquisition in their common ancestral lineage of a SUF-like minimal system Fe/S cluster pathway by lateral gene transfer. One of the isolates, *Skoliomonas litria*, appears to have lost all other known MRO pathways. No proteins were confidently assigned to the predicted MRO proteome of this organism suggesting that the organelle has been lost. The extreme mitochondrial reduction observed within this free-living anaerobic protistan clade demonstrates that mitochondrial functions may be completely lost even in free-living organisms.

Many unicellular eukaryote (protist) lineages have independently adapted to low-oxygen conditions. One of their most conspicuous adaptations are highly modified mitochondria referred to as “mitochondrion-related organelles” (MROs) that can function without oxygen. MROs have been best studied in the Metamonada, a supergroup of anaerobic protists with varied lifestyles including parasites, commensals, and free-living marine or freshwater flagellates and amoebae. All metamonad MROs studied to date are functionally reduced compared to aerobic mitochondria with their properties varying markedly across the group<sup>1</sup>.

The best studied metamonad MROs are hydrogenosomes that occur in the urogenital tract parasite *Trichomonas vaginalis*. The ‘hydrogenosomal’ ATP-producing pathway uses enzymes typically not found in aerobic mitochondria such as pyruvate:ferredoxin oxidoreductase (PFO) to oxidatively decarboxylate pyruvate to acetyl-CoA and reduce ferredoxin, an acetate:succinate CoA transferase that catalyzes the transfer of CoA onto succinate to produce succinyl-CoA, and an iron-only [FeFe]-hydrogenase that reduces protons to hydrogen gas by re-oxidizing the ferredoxin. ATP is produced from ADP and P<sub>i</sub> by the

<sup>1</sup>Institute for Comparative Genomics, Dalhousie University, Halifax, NS, Canada. <sup>2</sup>Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada. <sup>3</sup>Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden. <sup>4</sup>Department of Biology, Dalhousie University, Halifax, Canada. <sup>5</sup>Department of Biochemistry, Cambridge University, Cambridge, UK. <sup>6</sup>Section for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, Oslo, Norway. <sup>7</sup>Department of Biology, Lund University, Lund, Sweden. e-mail: [andrew.roger@dal.ca](mailto:andrew.roger@dal.ca)

Krebs cycle enzyme succinyl-CoA synthetase by substrate-level phosphorylation, converting succinyl-CoA to succinate in the process. Trichomonad hydrogenosomes also house a variety of other pathways including a mitochondrial-type iron-sulfur cluster (ISC) biogenesis system, amino acid metabolism, and oxygen detoxification<sup>2,3</sup>. Even more highly reduced MROs called ‘mitosomes’ are found in the metamonad gut parasite *Giardia intestinalis*. These organelles appear to lack any capacity to generate ATP; instead they participate in Fe/S cluster biogenesis via the mitochondrial-type ISC system<sup>4</sup>. Like trichomonads, *Giardia intestinalis* catalyzes acetyl-CoA production from pyruvate using PFO. However, this occurs in the cytoplasm. Acetyl-CoA is converted to acetate and ATP is produced by substrate-level phosphorylation catalyzed by the enzyme ADP-forming acetyl-CoA synthetase (ACS; EC 6.2.1.13)<sup>5</sup>. The most extreme form of MRO reduction in Metamonada can be found in the oxymonad gut commensal *Monocercomoides exilis* that has lost all traces of the organelle, making it the first truly amitochondriate eukaryote discovered<sup>6</sup>.

More recent investigations of newly discovered free-living metamonads have revealed novel configurations of MRO functions that do not fit into traditionally-defined organelle classes<sup>7</sup>. For example, the free-living metamonad *Dysnectes brevis* encodes MRO targeted [FeFe]-hydrogenases and associated proteins but lacks organelle-targeted enzymes for substrate-level phosphorylation, suggesting that while the MRO is capable of hydrogen metabolism, ATP is most likely produced in the cytosol<sup>1</sup>. A similar hydrogen-producing organelle and predicted cytosolic ATP production was reported in a transcriptomic investigation of the free-living metamonad *Barthelona* sp. PAPO20<sup>8</sup>. Recently, the enigmatic *Anaeramoeba* species—recently-discovered distant relatives to parabasalids like *Trichomonas*—were predicted to possess hydrogenosomes that have retained more mitochondrial systems than any other metamonad studied to date<sup>9</sup>. The MRO of the freshwater flagellate *Paratrimastix pyriformis* was recently investigated using spatial proteomic techniques which revealed that the organelle does not participate in either ATP production through extended glycolysis or Fe/S cluster production, and instead produces essential intermediates for the methionine cycle<sup>10</sup>. These studies of newly discovered free-living metamonads have provided important insights into the steps by which dramatically distinct MRO-types evolved within the Metamonada.

The multiple independent transitions from aerobically respiring mitochondria to MROs in the eukaryote tree in most cases have involved the complete loss of the mitochondrial genome itself, the loss of aerobic metabolism genes encoded in the nucleus, and the gain of genes encoding novel biochemical pathways by lateral gene transfer (LGT) from bacteria, archaea or other anaerobic protists<sup>11</sup>. For example, the ‘hydrogenosomal’ metabolism enzymes<sup>9,12,13</sup>, the ability to synthesize rhodoquinone<sup>14</sup> and oxygen defense enzymes<sup>15</sup> all appear to have been acquired by LGT in multiple distinct lineages from bacteria or other protists<sup>11,16</sup>. The most striking examples of the remodeling of mitochondrial functions in anaerobic protists involve the wholesale replacement of the highly-conserved essential mitochondrial ISC system by laterally acquired prokaryotic Fe/S cluster biogenesis systems. In the Archamoebae, a two-protein Fe/S biosynthesis pathway (recently classified as the minimal iron sulfur system (MIS) related to the nitrogen fixation system<sup>17</sup>) was acquired from bacteria and, in free-living members of the group, was subsequently duplicated into cytosolic and MRO-functioning versions<sup>18,19</sup>. A completely independent replacement of the mitochondrial ISC system occurred in the ancestors of the breviate flagellate *Pygsuia biforma* that acquired an archaeal-type SUF system<sup>20</sup>, recently classified as the SUF-like minimal system (SMS) pathway<sup>17</sup>. This breviate *SmsCB* gene (referred to as *SufCB* in ref. 20) in *Pygsuia biforma* was then duplicated to create cytosolic and MRO-targeted *SmsCB* proteins. In the oxymonads, yet another replacement of the ISC system occurred independently by the lateral acquisition of genes encoding a more complex multi-protein

SUF pathway<sup>6,21</sup>. This enabled the ancestral oxymonad to produce Fe/S clusters in the cytosol, easing the evolutionary pressure to keep the canonical ISC system, paving the way for the complete loss of MROs in *Monocercomoides*<sup>6</sup>. Phylogenetic analyses revealed that this SUF system is shared by a variety of preaxostylans and is likely bacterial in origin<sup>21</sup>.

To further investigate the patterns and process of genome and MRO evolution in metamonads, we sequenced the genomes of five recently-isolated free-living bacterivorous flagellates. Three of these organisms, the ‘skoliomonads’ *Skoliomonas litria*, *Skoliomonas* sp. GEM-RC (GEMRC) and *Skoliomonas* sp. TZLM3-RCL (RCL) were isolated from sediments of alkaline hypersaline soda lakes (for detailed descriptions see ref. 22; Fig. 1A–C). The other two are shallow marine-sediment dwelling ‘barthelonids’<sup>8</sup>: *Barthelona* sp. strain PCE (Fig. 1D) and *Barthelona* sp. PAPO20<sup>8</sup>. Unlike most previous studies of this kind that rely on incomplete transcriptomic data, we generated high quality contiguous genome assemblies, transcriptomes, and predicted proteomes from four of these five organisms. Our phylogenomic analyses demonstrate that these five metamonads form a highly supported clade, referred to as ‘BaSk’ (short for ‘Barthelonids+Skoliomonads’), that emerges as a sister group to Fornicata *sensu stricto* (diplomonads, retortamonads and *Carpediemonas*-like organisms (CLOs)). Genome size and coding capacity varies considerably across these organisms, with one possessing the smallest genome of a free-living eukaryote currently known. Their anaerobic metabolisms and MRO functional capacities also vary considerably. Unexpectedly, we found yet another instance of the horizontal acquisition of the *SmsCB* gene in their common ancestor. While several barthelonids possess both the *SmsCB* gene and core parts of a mitochondrial-type ISC system, all traces of the ISC system have been lost in the skoliomonads. For *Skoliomonas litria*, no MRO proteins could be confidently identified, suggesting it may lack the organelle altogether.

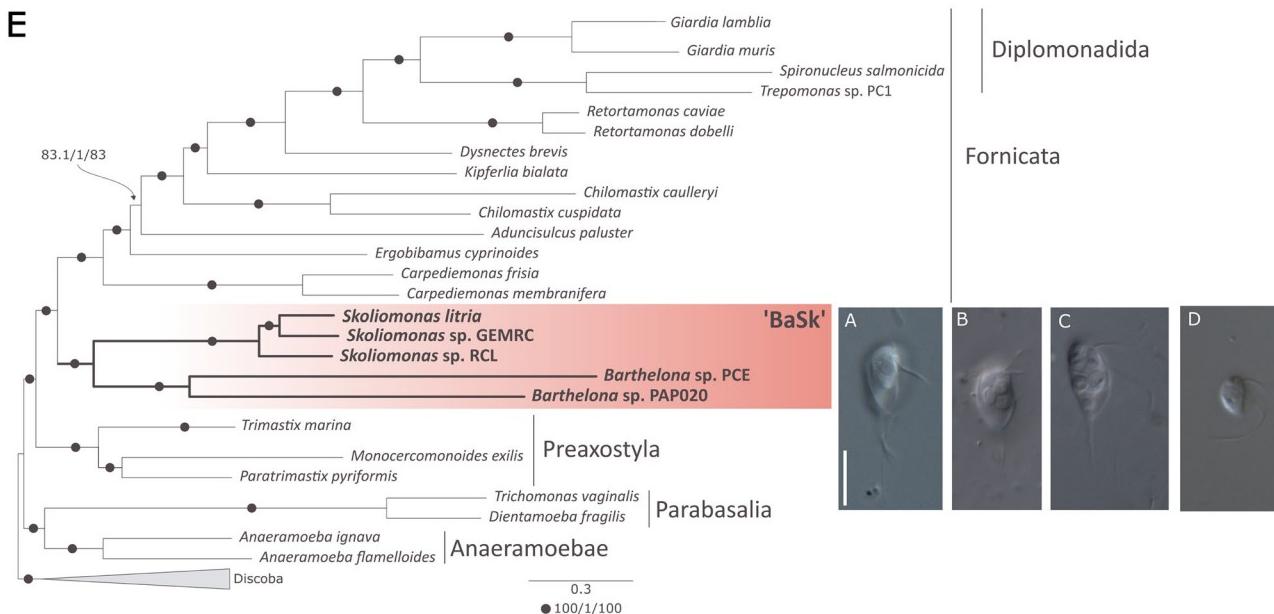
## Results

### Near-complete draft genomes of four novel metamonads

Near-complete genome assemblies of *Barthelona* sp. PCE, *Skoliomonas litria*, *Skoliomonas* sp. RCL and *Skoliomonas* sp. GEMRC were generated using a combination of Oxford Nanopore long-read sequencing for assembly and Illumina short-read sequencing for error-correction post-assembly. These assemblies were thoroughly decontaminated using the Eukfinder workflow<sup>23</sup>.

To assess the quality of these assemblies we used Merqury, a k-mer counting tool that uses genomic Illumina reads, assesses the completeness of the four de novo long-read assemblies<sup>24</sup>. Based on the Merqury analyses, all four assemblies exceeded 98% completion. Genome completeness was further supported by the fact that, for each of the four organisms, >99% of their respective eukaryotic-classified RNAseq reads mapped to the corresponding genome assembly. The combination of long and short-read data afforded assemblies that were high coverage and comprised of a few relatively large contigs (see Table 1). A draft genome assembly of *Barthelona* sp. PAPO20 was also generated using only Illumina reads (Table 1). This relatively fragmented assembly was included with other metamonads considered in this study to supplement our analyses concerning MRO reconstruction, metabolic pathway inventory, and phylogenomic inference, discussed below.

Gene predictions for these assemblies were generated by an in-house pipeline. To compare the qualities of the various gene predictions, BUSCO analyses were conducted based on the whole genomes themselves, the gene predictions and the transcriptome assemblies (Table 2). To assess potential improvement of the predicted protein set after curation, the gene predictions for *Skoliomonas litria* were further manually curated (Table 2). The percent of BUSCO genes identified amongst the gene predictions (i.e. predicted protein set) ranged from 63.7% (complete+fragmented) for *Barthelona* sp. PCE to



**Fig. 1 | 'BaSk' is a clade of anaerobic protists that branches as sister to all known Fornicata within the Metamonada. A–D** Differential interference contrast light micrographs of skoliomonad lineages *Skoliomonas litria* (A), *Skoliomonas* sp. GEMRC (B), and *Skoliomonas* sp. RCL (C) (imaging and methods described in ref. 22), as well as *Barthelona* sp. PCE (D) showing overall morphology. Scale bars indicate a length of 10  $\mu\text{m}$ . E A maximum likelihood (ML) phylogeny based on 174 concatenated aligned proteins encompassing 46,113 sites. The depicted topology was estimated using IQ-TREE under the LG + C60 + F +  $\Gamma$  model and was used as a guide tree for estimating the LG + PMSF(C60) + F +  $\Gamma$  model. Support under the

latter model was evaluated by SH-aLRT support percentage, aBayes support and nonparametric bootstrap percentage based on 100 nonparametric bootstrap replicates. The scale bar indicates the expected number of amino acid substitutions per site under the PMSF model. When at least one support value was less than maximal, all three are shown on the branch in the order SH-aLRT support percentage/aBayes/nonparametric bootstrap percentage. A black dot on the branch indicates all three approaches yielded full (i.e. 100/1/100) support. The BaSk clade is highlighted in red. See Supplementary Fig. 1 for full phylogenomic tree.

77.9% (complete + fragmented) for *Skoliomonas* sp. RCL. In general, the BUSCO scores of the gene predictions are very similar to those from the transcriptomes but the BUSCO genes of the former are less fragmented. The manual curation of *Skoliomonas litria* predicted proteins increased the reported BUSCO number by only one, indicating that the manual curation did not substantially improve the gene models. Note that these relatively low BUSCO scores are comparable to other well-characterized metamonad assemblies which tend to have highly divergent or missing orthologs relative to most other eukaryotes<sup>9,25</sup>.

The sizes of the draft genomes and the number of proteins predicted differ markedly between taxa (Table 1). Notably, at 10.2 Mbp, *Barthelona* sp. PCE appears to possess the smallest genome for a free-living eukaryote currently known. It is smaller than the compact genomes of the acidophilic alga *Galdieria sulphuraria* (13.7 Mbp, see ref. 26), *Saccharomyces cerevisiae* (12.07 Mbp, see ref. 27), and the genome of the smallest known free-living eukaryote, the marine alga *Ostreococcus tauri* (12.56 Mbp, see ref. 28). Genome reduction is well known to occur in parasitic and endosymbiotic microbes<sup>29,30</sup>, as well as in free-living microbes, where it is thought that compact genomes are an efficiency adaptation to nutrient-poor environments<sup>31</sup>. However, for heterotrophic phagotrophic protists, like *Barthelona* sp. PCE, nutrients are obtained predominantly from their bacterial prey, so it is unclear whether these generalizations hold and what other forces may drive genome reduction.

Two of the genome assemblies have apparent telomeric repeats on the ends of some contigs. *Barthelona* sp. PCE possesses two contigs with what appear to be telomeres at one end of the contig with an unusual telomere repeat sequence of 5'-TATATGGTCT-3'. Although this telomeric sequence is quite different from the eukaryotic 'consensus' repeat, other protists across the eukaryote tree of life have also been reported to have highly divergent telomeric repeats (see<sup>32</sup>); the significance of these unusual telomeres is unknown. The *Skoliomonas*

sp. GEMRC assembly has six telomere-to-telomere assembled chromosomes and an additional nine contigs with telomeres at only one end of the contig. The latter telomeres have the canonical 5'-TTAGGG-3' repeating sequences.

#### The BaSk taxa are the sister group to known members of Fornicata

To determine the phylogenetic affinities of the BaSks within the metamonads, we added orthologs from the five draft assemblies to a previously constructed phylogenomic data set of orthologous proteins<sup>33,34</sup>. After single-protein alignment manual curation to remove contaminants and paralogs, the final dataset of 174 highly conserved proteins was analyzed by maximum likelihood under the site-heterogeneous model LG + C60 + F +  $\Gamma$  (see Fig. 1E). The resulting tree confirms the placement of *Barthelona* sp. PAP020 as previously reported<sup>8</sup> and placed barthelonid strain *Barthelona* sp. PCE as its immediate sister. The skoliomonads form a clade with the barthelonid group, with *Skoliomonas litria* and *Skoliomonas* sp. GEMRC branching together to the exclusion of *Skoliomonas* sp. RCL. The 'BaSk clade' (Barthelonids + Skoliomonads) are sister to fornicates with maximum support (i.e. both the fornicate group and the fornicate + BaSk clade are supported with 100% non-parametric bootstrapping support, aBayes support, and approximate likelihood-ratio test support, all evaluated under the PMSF model).

Members of Fornicata possess a particular cytoskeletal feature, a distinct fibrous arch, that defines the group<sup>35</sup>. As there are no ultrastructural data for the flagellar apparatus cytoskeleton of any BaSk member to date, it is unclear whether they should be classified within the Fornicata or remain as a distinct sister group. Regardless, the BaSks 'deep' phylogenetic placement within the metamonads make them a key group for investigations into the origins and evolutionary trajectories of genes involved in parasitism and MRO reduction within the Metamonada supergroup.

**Table 1 | Summary statistics for the genomes, transcriptomes, and predicted proteomes of the BaSk clade**

	<i>Skoliomonas litria</i>	<i>Skoliomonas</i> sp. GEMRC	<i>Skoliomonas</i> sp. RCL	<i>Barthelona</i> sp. PCE	<i>Barthelona</i> sp. PAPO20
<sup>a</sup> Genome Coverage (Nanopore)	41x	119x	34x	66x	Illumina only
<sup>a</sup> Genome Coverage (Illumina)	206x	116x	93x	308x	29x
Number of Illumina Reads Generated	53.7 Million	26.8 Million	85.9 Million	75.4 Million	70.0 Million
% RNAseq reads mapping	99.4	99.9	99.6	99.3	-
Mercury Completeness (%)	98.7	98.6	98.4	98.4	-
Genome Size	31.9 Mbp	33.5 Mbp	21.1 Mbp	10.2 Mbp	13.2 Mbp
Contigs (Number)	16	21	18	8	1035
N50	3,724,615	2,551,763	2,069,101	1,587,787	24,218
L50	4	6	4	3	156
% GC	35.8	36.9	36.0	40.3	30.5
%x Repetitive	55.4	48.0	28.1	4.46	-
Predicted Proteins (Number)	15,224	14,268	9198	5064	6305

<sup>a</sup>Median contig coverage.**Table 2 | BUSCO scores for members of the BaSk clade**

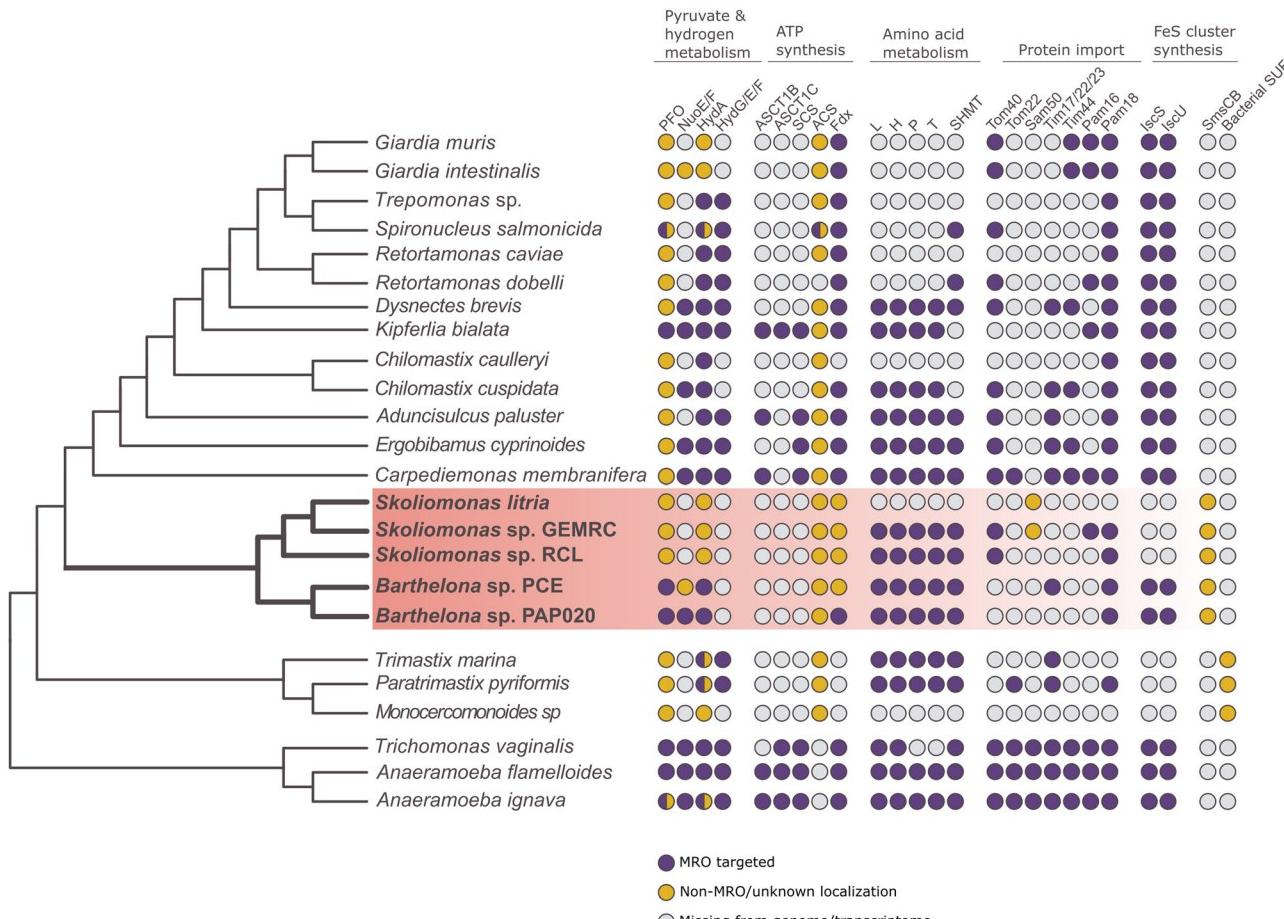
	Complete single copy	Complete duplicated	Total complete	Fragmented	Missing
<i>S. litria</i> - Genome	178	4	182 (60.1%)	15 (4.9%)	106 (35.0%)
<i>S. litria</i> - Prediction	215	9	224 (73.9%)	7 (2.3%)	72 (23.8%)
<i>S. litria</i> - Transcriptome	190	24	214 (70.6%)	16 (5.3%)	73 (24.1%)
<i>S. litria</i> - Curated Predictions	208	17	225 (74.3%)	7 (2.3%)	71 (23.4%)
<i>S. sp.</i> GEMRC - Genome	168	9	177 (58.4%)	27 (8.9%)	99 (32.7%)
<i>S. sp.</i> GEMRC - Prediction	183	14	197 (65.0%)	29 (9.6%)	77 (25.4%)
<i>S. sp.</i> GEMRC - Transcriptome	139	77	216 (71.3%)	9 (3.0%)	78 (25.7%)
<i>S. sp.</i> RCL - Genome	182	4	186 (61.4%)	16 (5.3%)	101 (33.3%)
<i>S. sp.</i> RCL - Prediction	212	10	222 (73.3%)	14 (4.6%)	67 (22.1%)
<i>S. sp.</i> RCL - Transcriptome	193	27	220 (72.6%)	14 (4.6%)	69 (22.8%)
<i>B. sp.</i> PCE - Genome	142	3	145 (47.9%)	15 (4.9%)	143 (47.2%)
<i>B. sp.</i> PCE - Prediction	161	7	168 (55.4%)	25 (8.3%)	110 (36.3%)
<i>B. sp.</i> PCE - Transcriptome	157	5	162 (53.5%)	31 (10.2%)	110 (36.3%)

### BaSks all possess highly reduced MROs and share the same core anaerobic metabolism

We searched the genomes, transcriptomes, and predicted protein sets of all four BaSks using a variety of bioinformatic approaches (see Methods) to recover proteins related to MRO function and anaerobic metabolism. We also utilized a draft short-read Illumina genome assembly of *Barthelona* sp. PAPO20 in these searches to supplement our comparisons. Several mitochondrial targeting sequence prediction programs were used to estimate the probability of mitochondrial localization of each predicted protein in each of the five genomes. We used these predicted proteins as queries against the nr database using BLAST and looked for outputs relating to potential mitochondrial-localizing proteins. The predicted proteins from the BaSk genomes were also used as queries in a BLAST search of a database comprised of mitochondrial proteomes from six diverse eukaryote lineages and the best hits from each query were retained and inspected. Next, we utilized datasets of known metamonad MRO proteins to query the predicted protein sets from the BaSk genomes, as well as directly investigate the genomes and transcriptomes. This was done through both BLAST search and HMM profile searching. In cases where search results were ambiguous, we utilized appropriate Pfam profiles to search for protein domain information. Finally, we assessed candidate proteins by constructing phylogenetic trees of corresponding MRO proteins with additional cytosolic homologs, to further verify the probable localization of BaSk proteins. A summary of the results of all

of these searches can be found as Fig. 2. Below we discuss the systems we investigated in detail by first highlighting the pathways that, when present, function exclusively in MROs and are therefore strong indicators of the presence of the organelle.

We first investigated whether the BaSk genomes encoded homologs of the glycine cleavage system (GCS), an exclusively mitochondrial pathway which converts glycine to ammonia and carbon dioxide while also reducing NAD<sup>+</sup> to NADH and generating 5,10-methylene-tetrahydrofolate (5,10-CH<sub>2</sub>-THF)<sup>36</sup>. This pathway is commonly retained in MROs of free-living anaerobic protists including some metamonads<sup>1</sup>, but is sometimes missing in parasitic lineages (reviewed in ref. 11). Except for *Skoliomonas litria*, all BaSks encode the complete GCS (i.e. the T-, P-, L- and H-proteins). Additionally, all BaSks possessing the GCS also have a single serine hydroxymethyltransferase (SHMT) gene. This enzyme produces serine using both glycine and the 5,10-CH<sub>2</sub>-THF produced by the GCS. GCS and SHMT are found in the MROs of a variety of diverse free-living metamonads including *Paratrimastix pyriformis*<sup>10,37</sup> and all of the non-diplomonad fornicates (i.e. CLOs and *Chilomastix cuspidata*; see ref. 1). Complete absence of the GCS in metamonads is seen in the parasitic diplomonads *Giardia intestinalis* and *Spironucleus salmonicida*<sup>1,38</sup>. Other parasitic metamonads such as *Trichomonas vaginalis* retain some of the GCS subunits<sup>39</sup>, but have repurposed them to function in a peroxide detoxification pathway<sup>40</sup>. It was particularly unexpected, therefore, that we found neither genes encoding GCS nor SHMT in *Skoliomonas litria*, making it the only known free-living



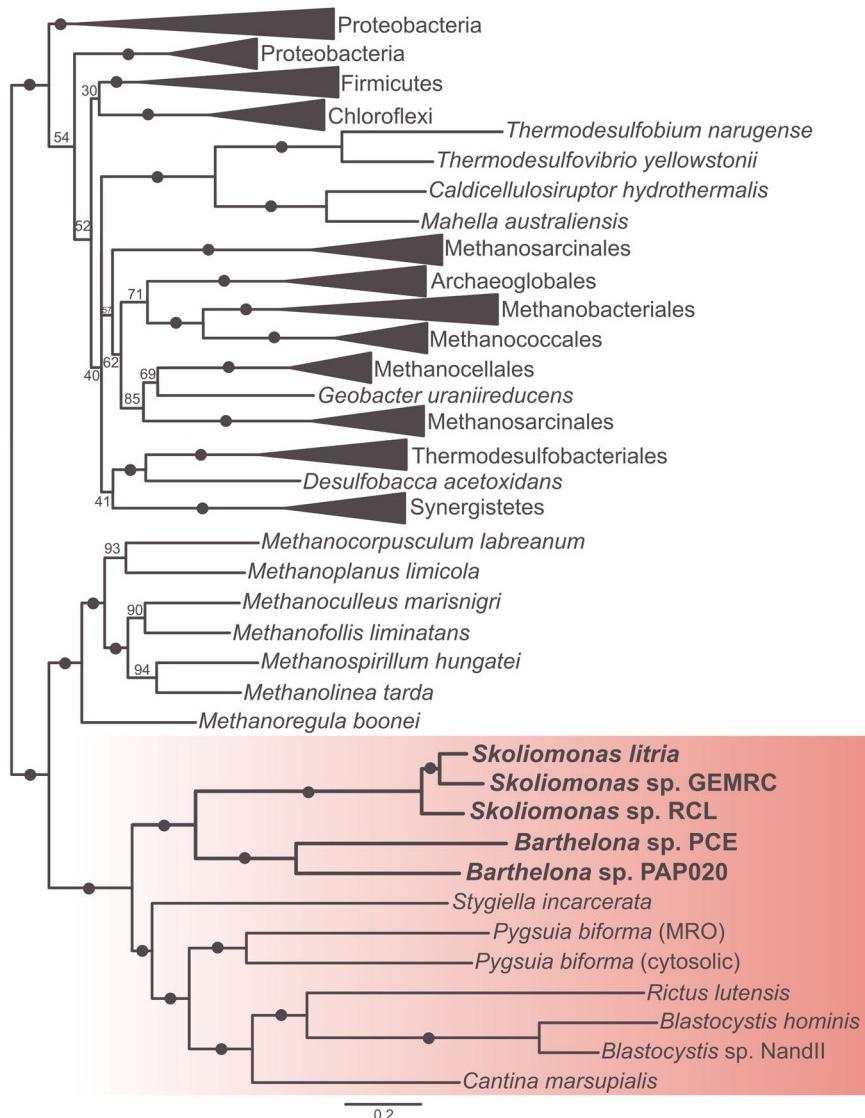
**Fig. 2 | Presence, absence and predicted localization of key MRO and energy metabolism enzymes in ‘BaSks’ and other metamonads.** A schematic phylogeny of the Metamonada is shown on the left based on Fig. 1, with the BaSk clade highlighted in red. The pathways depicted include anaerobic extended glycolysis/energy metabolism enzymes, conserved MRO proteins, MRO protein targeting machinery components, and the SmsCB fusion protein. Coloured versus grey circles indicate presence/absence of the protein from predicted proteomes and/or transcriptome data and the colours indicate probable localization. Circles split in half indicate that more than one paralog was present with different predicted localizations. Protein abbreviations: PFO Pyruvate:ferredoxin oxidoreductase, NuoE/F Respiratory-chain NADH dehydrogenase 24/51 kDa subunit, HydA Iron-only hydrogenase, ASCT1B Acetate:succinyl-CoA transferase B, ASCT1C

Acetate:succinyl-CoA transferase C, SCS Succinyl-CoA synthase, ACS Acetyl-CoA synthase, Fdx Ferredoxin, L Glycine cleavage system subunit L, H Glycine cleavage system subunit H, P Glycine cleavage system subunit P, T Glycine cleavage system subunit T, SHMT Serine hydroxymethyltransferase, Tom40 Translocase of outer mitochondrial membrane 40, Tom22 Translocase of outer mitochondrial membrane 22, Sam50 Sorting and assembly machinery 50, Tim17/22/23 Translocase of the inner membrane 17/22/23, Tim44 Translocase of the inner membrane 44, Pam16 Presequence translocase-associated motor 16, Pam18 Presequence translocase-associated motor 18, IscS Iron-sulfur cluster assembly enzyme cysteine desulfurase, IscU Iron-sulfur cluster assembly enzyme scaffold, SufCB Sulfur formation CB fusion protein.

metamonad completely lacking this pathway, and the only BaSk member to have no detectable MRO amino acid metabolism. The MROs within the BaSks are therefore key to understanding the reduction of amino acid metabolism that is usually only ever seen in parasitic metamonads. As previously reported<sup>8</sup>, *Barthelona* sp. PAP020 contains a OsmC homolog, but we were unable to detect this enzyme in any of the other BaSk data, including the *Barthelona* sp. PAP020 genome. Because this enzyme lacks a mitochondrial targeting sequence (MTS), it is unclear where this protein localizes within the *Barthelona* sp. PAP020 cell, and if it participates in peroxide detoxification in the *Barthelona* sp. PAP020 MRO, as it does in the MRO of *Trichomonas*<sup>40</sup>.

Next, we investigated the highly conserved mitochondrial Fe/S cluster (ISC) pathway that is also exclusively found within mitochondria or MROs. Unexpectedly, amongst all the BaSks, only the barthelonid PCE and PAP020 assemblies encoded any detectable ISC system components. These two genomes encode cysteine desulfurase NFS1 (also known as ISCS), the scaffold protein ISCU and a ferredoxin. Notably, they also encode Fe/S cluster containing NuoE and NuoF subunits of complex I, which, like core ISC components, are homologs

which are only known to function in mitochondria and the MROs of other organisms. Surprisingly, with the exceptions of mitochondrial Hsp70 (discussed further below) and ferredoxin, no other genes encoding ISC pathway components were detected in the barthelonids or any other of the BaSks (see Supplementary Data 1 for ISC proteins queried). The absence of some of these proteins, like Isd11 and ferredoxin reductase (Arh1), is not unexpected as they have not been detected in most metamonads<sup>41</sup>. The absence of other early ISC pathway proteins implicated in [2Fe-2S] cluster synthesis (e.g. frataxin, Jac1 (HscB) and Mge1 (GrpE)) is more puzzling as these components are usually detected in metamonads<sup>41</sup>. Similarly, some late ISC pathway components implicated in [4Fe-4S] cluster synthesis<sup>42</sup> are found in other metamonads (e.g. Grx5, Isa2, NFU1 and BolA<sup>41</sup>), but were not detected in any BaSks (Supplementary Data 1). This is strange given that the barthelonids possess mitochondrial NuoF subunits that typically house an [4Fe-4S] cluster<sup>43</sup>. This means that if these NuoF subunits are indeed found within the MROs of the barthelonids, as they are with other eukaryotes, their [4Fe-4S] clusters are being produced by an unknown mechanism.



**Fig. 3 | A phylogeny of the SmsCB fusion protein in a variety of anaerobic eukaryotes (red) and their closest bacterial and archaeal homologs.** The alignment contains the SmsCB fusion protein from the eukaryotes and concatenated

SmsC and SmsB proteins from the prokaryotes. This ML tree was estimated under the LG + C60 + F + Γ model using IQ-TREE. Ultrafast bootstrap (UFBOOT2) values are displayed on the branches, with dots indicating UFBOOT2 values over 95%.

Despite the lack of ISC system proteins in most BaSks, all BaSk members appear to encode a similar partial cytosolic CIA system that includes apoprotein targeting components CIA1, CIA2, and NARI, the NBP35 scaffold protein, and TAH18, which aids in maturation of Fe/S cluster proteins. This reduced CIA pathway is not unlike the CIA pathway found within *Giardia intestinalis*<sup>44</sup>.

#### BaSks possess simple archaeal-type SUF-like minimal system (SMS) fusion proteins

We investigated the possibility that the lack of ISC system components in BaSk members may be complemented by the presence of an alternative Fe/S cluster system such as the minimal Fe/S (MIS) or sulfur mobilization (SUF) systems found in a number of other anaerobic protists<sup>17,20,21,45,46</sup>. In all of the BaSk genomes, we were able to identify a simple archaeal-type SUF-like minimal system (SMS) consisting of a gene encoding an SmsCB fusion protein that was most similar to the simple archaeal-type SmsCB systems found in the breviate *Pygsuia biforma*<sup>20</sup>, the anaerobic jakobid *Stygiella incarcerata*<sup>13</sup> and gut commensal opilinatan stramenopile *Blastocystis*<sup>47,48</sup>. To investigate the origin of this protein, we performed phylogenetic analysis separately for the SmsC and SmsB domains (Supplementary Figs. 2 and 3) before

concatenating the domains together for the final tree (Fig. 3). The BaSk SmsCB proteins form a strongly supported clade that is sister group to all archaeal-type SmsCB fusion proteins from anaerobic protists. Consistent with previous reports, the anaerobic protistan clade of SmsCB fusion proteins is sister group to the orthologs from members of the Methanomicrobiales order of Archaea<sup>48</sup>. This phylogenetic pattern is most easily explained if there were an original transfer and fusion of SmsCB to an anaerobic protist from a methanomicrobiales donor, followed by a series of eukaryote-to-eukaryote LGT events amongst disparate lineages of anaerobic eukaryote lineages. The deeply branching position of the BaSk clade amongst the anaerobic protists suggests it is possible that an ancestral BaSk was the first recipient of these genes and an early offshoot from the BaSk lineage passed the fusion gene to other anaerobic protists, although it is impossible to rule out alternative scenarios. Regardless, this SMS system is phylogenetically distinct, and has a separate origin from, the more complex SUF system found within the metamonad preaxostylans (i.e. oxymonads and *Paratrimastix pyriformis*)<sup>21</sup> and therefore represents an interesting case of convergence by LGT within the Metamonada. The resulting reduction and loss of the ISC system following acquisition of a SMS system is also similar between these two lineages

and the breviate *Pygsuia biforma*. The gain of an alternate Fe/S cluster synthesis system in these cases likely ‘preadapts’ the ancestral organism to the loss of the ISC system in some of the lineages (e.g. an ancestor of *Skoliomonas* sp. GEMRC, *Skoliomonas* sp. RCL and *Skoliomonas litria*). Unlike *Pygsuia biforma* and the preaxostylans, *Barthelona* sp. PCE and *Barthelona* sp. PAP020 retain a few core components of the early ISC system in addition to possessing the novel SMS system, suggesting that their MRO ISC systems are at least partially functional. This retention of the ISC proteins may be related to the possession of Fe/S cluster-containing proteins such as the NuoE and NuoF complex I proteins that are predicted to function in their MROs. Although we cannot definitively determine the subcellular localization of the SmsCB in BaSks bioinformatically, it is likely that this system is cytosolic as in other anaerobic protists such as *Blastocystis* spp<sup>48</sup>. If it is indeed cytosolic, it is unclear whether it would function in coordination with the cytosolic CIA Fe/S biogenesis system. One possibility is that the CIA system and the SmsCB system may function in parallel with each responsible for the assembly of Fe/S clusters for different specific subpopulations of apoproteins. Alternatively, the two systems may work together, similar to how the ISC system coordinates with the CIA pathway<sup>44</sup>. In the latter scenario, the SmsCB fusion protein would provide the CIA pathway with a sulfur-containing intermediate, which the CIA pathway would then use to assemble Fe/S clusters and insert them into recipient apoproteins.

In any case, the skoliomonads completely lack any dedicated ISC components and do not possess proteins with identifiable cysteine desulfurase domains as assessed by HMM profile searches and phylogenetic analysis (see Supplementary Fig. 4). This is surprising because for most Fe/S cluster synthesis systems, cysteine is the source of sulfur and a cysteine desulfurase is the key enzyme involved in mobilizing it. We suspect that the SmsCB system of these organisms may directly utilize sulfide from the environment as the sulfur source for Fe/S clusters, as previously suggested for *Pygsuia biforma*<sup>20</sup>. Indeed, direct sulfide utilization has been demonstrated for methanogenic archaea that possess a simple cysteine-desulfurase-lacking SMS pathway and which also live in sulfidic conditions<sup>49</sup>. The exact mechanism by which such organisms construct Fe/S clusters is unknown and more work needs to be done to understand their biogenesis in these organisms.

#### Presence or absence of other conserved mitochondrial proteins in the BaSk taxa

The next highly conserved mitochondrion/MRO proteins that we investigated were the mitochondrial orthologs of the molecular chaperones chaperonin-60 (Cpn60) and hsp70. To distinguish cytosolic Chaperonin Containing TCP-1 (CCT) subunits from mitochondrial Cpn60 (GroEL) homologs and cytosolic and endoplasmic reticulum (ER) paralogs of hsp70 from their mitochondrial Hsp70 (i.e. DnaK orthologs), we used phylogenetic analyses. For *Skoliomonas litria* and *Skoliomonas* sp. RCL, all Cpn60 candidates identified distinctly branched with CCT subunit clades (Supplementary Fig. 5); no bona fide mitochondrial Cpn60s could be identified from these species. *Skoliomonas* sp. GEMRC, *Barthelona* sp. PCE, and *Barthelona* sp. PAP020 appear to have both CCTs and mitochondrial Cpn60. Phylogenetic analysis of Hsp70 proved more difficult, due in part to a greater degree of similarity between paralogous copies of the chaperone. To supplement this analysis, signature insertion-deletions in the multiple alignments were identified that were used in conjunction with well-annotated Hsp70 yeast homologs<sup>50</sup> to group sequences into cytosolic/ER versus mitochondrial type Hsp70 groups (see Supplementary Fig. 6). This information was then checked against the phylogeny for subcellular localization predictions of BaSk Hsp70 candidates. From this analysis, mitochondrial-type Hsp70s were identified in all BaSks except *Skoliomonas litria*. Within MROs, Cpn60 and Hsp70 aid in protein translocation into the matrix and help newly translocated

proteins adopt their native structure. These chaperones are ubiquitously predicted to function in a variety of MROs, so it is notable that they are missing in some of the BaSks data. For *Skoliomonas litria*, we did not detect any genes encoding MRO chaperones, while *Skoliomonas* sp. RCL appears to contain Hsp70 but lacks Cpn60. This further highlights the highly reduced nature of the MROs found within these protists, especially within *Skoliomonas litria*.

All mitochondria and MROs have membrane associated proteins, including the TIM/TOM complex proteins involved in protein translocation into the organelle. To search for these, we used HMM profile searching because many of these sequences are highly divergent and not detectable by pairwise alignment-based search algorithms. The main channel forming protein in the TOM complex, Tom40, was only identifiable in *Skoliomonas* sp. RCL and *Skoliomonas* sp. GEMRC, while the receptor Tom70 was identified in *Barthelona* sp. PCE, *Barthelona* sp. PAP020, and *Skoliomonas* sp. RCL. TIM subunits were generally absent except for Tim17 that was identified in *Barthelona* sp. PCE only. Except for *Skoliomonas litria*, all BaSk members appeared to have the co-chaperone Pam18 subunit of the presequence translocase-associated motor (PAM) complex in addition to the core mtHsp70 subunit. Pam16 was only identified in *Skoliomonas* sp. GEMRC and Pam17 appeared to be absent.

In mitochondria, mitochondrial processing peptidases (MPPs) and presequence proteases (PreP) remove the MTS after protein import into the organelles<sup>51</sup>. We identified MPP/PreP homologs only in *Skoliomonas litria*, *Skoliomonas* sp. GEMRC, and *Skoliomonas* sp. RCL and none had predicted MTSSs. We further investigated their domain structures and phylogenetic position amongst MPP/PreP-related protein families in eukaryotes and prokaryotes. For our phylogenetic analysis we used a recently published multiple alignment of all major families of M16 proteases<sup>51</sup> supplemented with top 20 BLAST hits in the nr database to each of the *Skoliomonas litria* sequences. Our phylogenetic analyses (Supplementary Fig. 7A) showed that skoliomonad sequences branch in multiple different positions in the tree, but none are closely related to eukaryote mitochondrial MPPs or PrePs. All but one of these proteins have a four-domain structure that is characteristic of PreP and SPPs but distinct from the two domain structure of the MPP family<sup>51</sup>. One *Skoliomonas* sp. RCL sequence has two domains that appear to have evolved separately from the two-domain MPP group (Supplementary Fig. 7B). Notably, the two *Skoliomonas litria* M16 homologs branch with skoliomonad orthologs in two distinct clades. One of these branches from within a group of entirely bacterial PreP or MAG peptidase sequences and the other is a sister group of uncharacterized homologs from invertebrates (Metazoa) which lack predicted MTSSs (except for one homolog in *Branchiostoma* that is the result of a recent duplication). This analysis suggests that none of the M16-related proteases in skoliomonads are orthologs of mitochondrial MPP or PreP; they are likely M16 proteases with separate cytosolic functions.

A candidate mitochondrial carrier protein (MCP) was identified in *Skoliomonas litria* but its substrate specificity could not be determined by phylogenetic analysis (Supplementary Fig. 8). Whether this protein functions in the MRO membrane is unclear as MCPs are also known to function in peroxisomes or other subcellular organelles<sup>52</sup>. Finally, a candidate homolog of Sam50 was identified in *Skoliomonas litria* and *Skoliomonas* sp. GEMRC. However, in phylogenetic analyses these proteins did not emerge within the clade of mitochondrial Sam50 orthologs, but instead formed an independent group emerging halfway between mitochondrial Sam50 and bacterial BamA clades (Supplementary Fig. 9A). Orthologs of this Sam50/BamA-like protein were not detectable in any of the other BaSk genomes, even when this sequence was included in the HMM profile used for searching. To determine if this Sam50/BamA-like protein folded into a structure resembling the mitochondrial orthologs, we used AlphaFold 2.0<sup>53</sup> to predict its structure. The resulting structures (Supplementary

Fig. 9B, C) show that while these candidate Sam50 proteins possess what could be a intermembrane POTRA domain, the beta-barrels are smaller and adopt a different confirmation compared to Sam50 homologs of yeast (Supplementary Fig. 10) and other well characterized examples<sup>54</sup>.

Among the five MRO predicted proteomes we present here, *Skoliomonas* sp. GEMRC and *Skoliomonas* sp. RCL contain the most candidate MRO translocon proteins, which is similar to the repertoire found in *Giardia intestinalis*<sup>1</sup> with the notable exception of the absence of GrpE in all members of the BaSk clade. *Barthelona* sp. PCE, *Barthelona* sp. PAP020 and *Skoliomonas litria* possess a remarkably reduced MRO translocon complex by comparison. This may be due to the highly divergent nature of membrane bound MRO proteins, which makes detecting such proteins challenging<sup>55</sup>. Alternatively, in the case of *Skoliomonas litria* which lacks all the highly conserved GCS, ISC, protein import/refolding proteins of MROs, it is unclear if the Sam50/BamA-like protein, the MCP and MPP/M16 metalloprotease homologs are targeted to an organelle at all. It is possible that *Skoliomonas litria*, like the distantly-related *Monocercomonoides exilis*, might completely lack a mitochondrial compartment and these proteins have taken on roles elsewhere in the cell. Further work is needed to determine the cellular localization of these proteins and, if it exists, the function of *Skoliomonas litria*'s MRO in the absence of all known conserved mitochondrial pathways and systems.

MTS prediction is a known challenge in inferring MRO function in anaerobic protists, including metamonads<sup>2,38,56</sup>, due to highly divergent or missing MTSs in these organisms. It has been suggested that the lack of a long positively charged MTS in metamonads is due to the loss of the proton gradient in the organelles<sup>57</sup>, though recent analysis of canonical MTSs found in the Anaeramoebae – deep-branching metamonads – suggest that this may not always be the case<sup>9</sup>. Previous work has shown that some MRO proteins rely on cryptic internal signals to target proteins to the organelle with high fidelity<sup>57,58</sup>. Furthermore, most available MTS prediction software tools were trained on experimentally validated datasets of targeted proteins that are lacking for most protists. For these reasons, making accurate MTS predictions in metamonads is exceedingly difficult. Here, we employed several different MTS prediction software tools to find and annotate these signals in BaSks (see Supplementary Data 1). We found several examples of false positives in *Skoliomonas litria* and *Skoliomonas* sp. GEMRC, as homologs of transposons and nuclear targeting proteins were predicted to possess an MTS by some targeting prediction tools with high confidence. Because of this, we considered proteins with MTS predictions from two or more MTS prediction software to be “strongly supported” targeting signals. Based on this criterion, *Skoliomonas litria* lacked any strongly supported targeting signals, while strongly supported targeting signals were rare in all other BaSk members. *Barthelona* sp. PCE contains strongly supported MTSs predicted on three proteins: an [FeFe]-hydrogenase, the T-protein of the glycine cleavage system (GCST), and on the IscU scaffold protein. *Barthelona* sp. PAP020 also contained a strongly supported MTS on its ortholog of IscU, as well as a different component of the glycine cleavage system (GCSP2). Another MTS detected was found on pre-sequence translocated-associated motor protein PAM18 in *Skoliomonas* sp. GEMRC. Interestingly, in *Skoliomonas* sp. RCL we find that a cytosolic CCT-gamma subunit contains a strongly predicted MTS, where in other BaSks, all cytosolic CCT subunits appear to be missing an MTS. However, when aligned to other BaSk CCT-gamma subunits, it does not appear to have an obvious N-terminal extension. As *Skoliomonas* sp. RCL has no mitochondrial Cpn60 homolog but does contain proteins which are homologous to MRO-localized proteins in other metamonads, it is possible (but unprecedented) that this CCT subunit has replaced the mitochondrial homolog in this species by acquiring an MTS. *Skoliomonas* sp. RCL does not have an additional

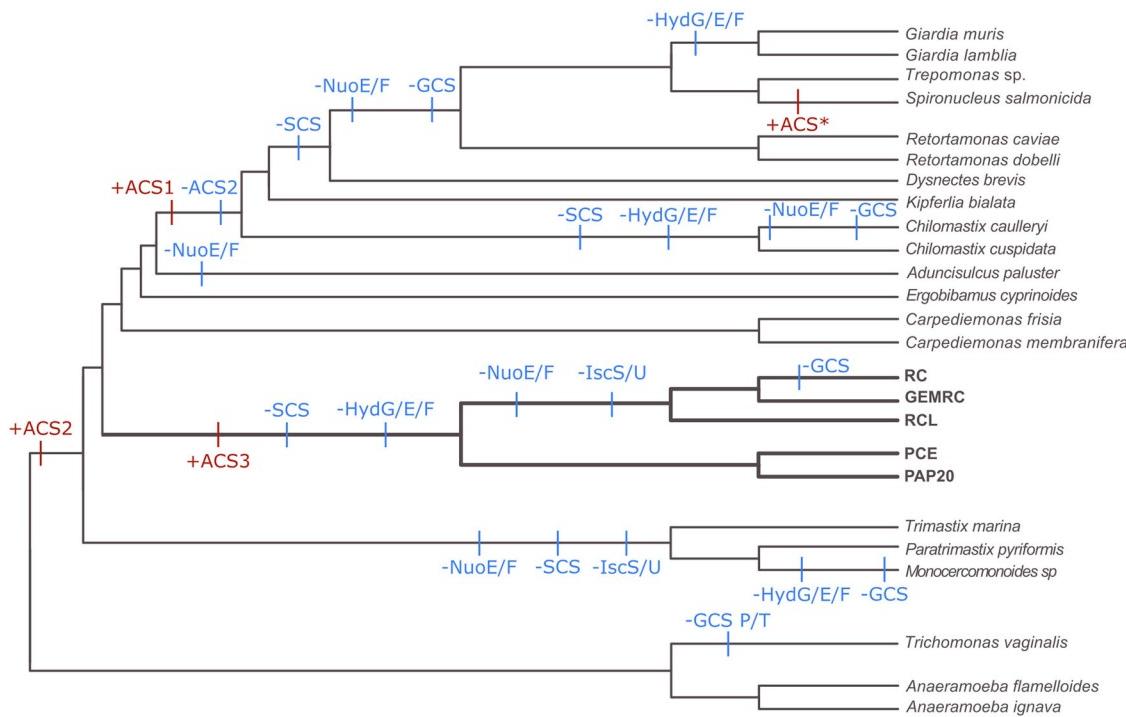
copy of the gamma subunit of CCT. Follow up localization experiments are needed to deduce the subcellular localization of this particular CCT subunit.

### The BaSks share a core ATP generation pathway

All BaSk members share a similar hypothetical ATP production pathway that is reminiscent of the pathway found in *Giardia intestinalis*<sup>4,59</sup>. In this pathway, pyruvate:ferredoxin oxidoreductase (PFO) catalyzes the oxidation of pyruvate in the presence of ferredoxin and CoA to produce acetyl-CoA, CO<sub>2</sub> and reduced ferredoxin. Acetyl-CoA is then transformed to acetate by ACS and, in the process, converts ADP to ATP. Ferredoxin is reoxidized by [FeFe]-hydrogenase (HydA) that passes electrons to protons creating H<sub>2</sub> gas. Notably, pyruvate-formate lyase (PFL) was also detected in all BaSk members, allowing an alternative pathway of acetyl-CoA production from pyruvate<sup>12</sup> (see Supplementary Fig. 11). Note that there are multiple copies of both HydA and PFO in each BaSk member, some of which possess additional domains encoding ferredoxin or flavodoxin-like proteins (see Supplementary Data 2). No components of typical ATP production pathways found in trichomonad hydrogenosomes, such as acetate:succinyl-CoA transferase or succinyl-CoA synthase<sup>60,61</sup>, were detected in any of the BaSks. Though [FeFe]-hydrogenase is present in each BaSk member, none of the [FeFe]-hydrogenase maturase proteins (HydG, E or F) were detected. This is notable because these maturases, when present, are always targeted to mitochondria or MROs. This, combined with the lack of MTS detected in any of the proteins from the above pathway, leads us to assume that ATP and H<sub>2</sub> gas are likely produced in the cytosol of the skoliomonads.

However, the pathways in the barthelonids appear to be slightly different in function and likely localization. Both *Barthelona* sp. PCE and *Barthelona* sp. PAP020 encode NuoE and NuoF NADH-dehydrogenase subunits of complex I that are known to function in the mitochondrial/MRO matrix in eukaryotes. It is widely thought that in the anaerobic protists that possess them, NuoE and NuoF likely form a complex with an MRO-localized HydA to perform simultaneous NADH and ferredoxin (Fd<sup>-</sup>) oxidation by an electron confurcating reaction that produces H<sub>2</sub> gas, NAD<sup>+</sup> and Fd<sup>3,62</sup>. For this reason, and the fact that one of the HydA proteins of *Barthelona* sp. PCE and *Barthelona* sp. PAP020 have a predicted MTS, we infer that both barthelonids likely have MRO-associated hydrogenase activity, as recently suggested for *Barthelona* sp. PAP020<sup>8</sup>. However, unlike the hypothesis of Yazaki and colleagues (see Fig. 4b in<sup>8</sup>), we suggest that the reduced ferredoxin may, in fact, get produced within their MROs by PFO activity. If so, it is also possible that at least one of the two ACS homologs in each of these organisms is also MRO-localized and produces ATP within the organelle, though no MTSs are predicted on any of the ACS homologs. Further localization experiments in these barthelonids are needed to test these hypotheses.

Previous studies have shown that some of the enzymes involved in ATP synthesis found within metamonads have a complex phylogenetic history. In particular, ACS is thought to have been transferred into the metamonads multiple times through LGT<sup>1</sup>. To determine the evolutionary histories of the BaSk ACSs, we constructed a phylogenetic tree containing ACSs from a variety of metamonads, which can be found as Supplementary Fig. 12. Our results show that, as in previous studies, most metamonad ACSs branch in one of two clades, known as ACS1 and ACS2. The ACS homologs found in the BaSks also form two distinct groups; one appears to be related to ACS2 found within *Carpediemonas membranifera*, *Ergobibamus cyprinoides*, and *Trimastix*. Additional ACS2 homologs from *Monocercomonoides* and *Paratrimastix pyriformis* were also identified, which also branch with the metamonad ACS2 group. This suggests that ACS2 was inherited in the last common ancestor of Pre-axostyla and the fornicates + BaSk. The BaSk also possess a second ACS homolog which branches separately from other metamonad ACSs. This enzyme seems to be most closely related to ACS found within *Streblomastix*, *Monocercomonoides* and *Blastocystis*, and is distinct from other types of ACS in the tree. As previously reported in an analysis of



**Fig. 4 | Gain and loss of anaerobic enzymes and MRO proteins over the tree of Metamonada.** Gains (red) and losses (blue) of key metabolic pathway components discussed throughout this paper are mapped on a schematic phylogeny of Metamonada (based on Fig. 1). Protein abbreviations are: HydG/E/F Iron hydrogenase

maturase G/E/F, ACS Acetyl-CoA synthase, NuoE/F Respiratory-chain NADH dehydrogenase 24/51 kDa subunit, GCS Glycine cleavage system, SCS Succinyl-CoA synthase, IscS/U Iron-sulfur cluster assembly enzyme system. The *Spironucleus* ACS is denoted as ACS\* as an ACS “type” could not be confidently assigned.

the *Barthelona* sp. PAPO20 transcriptome<sup>8</sup>, this is likely to be a novel homolog of ACS (ACS3) and was probably transferred into metamonads through a LGT event distinct from the event that gave rise to ACS1 and ACS2. Our analysis also suggests that the ACS found within *Spironucleus salmonicida* MROs may not be ACS2, as previously reported<sup>1</sup> because, in our analyses, the *S. salmonicida* ortholog (denoted ACS\* in Fig. 4 and Supplementary Fig. 12) did not form a clade with other metamonad ACS2 sequences but instead grouped with several TACK archaea, although the support values for this relationship are low. A diagram summarizing the gains and losses of these proteins, along with other key metabolic processes and proteins mentioned above, can be found as Fig. 4. These patterns suggest that there have been at least three, and possibly four, independent transfers of ACS into the metamonads. These transfers occurred at very different points in time along the evolution of various metamonad groups and highlights the role of LGT in the evolution of anaerobic metabolism.

To investigate the expression level of these key metabolic enzymes, we mapped RNASeq reads to each gene model, then normalized read counts and compared the fragments per kilobase of transcript per million mapped reads (FPKM) value amongst these gene models (see Supplementary Data 3). We found that some enzymes relating to anaerobic metabolism were amongst the top 100 most highly expressed genes in all four of the BaSk queried, though the specific anaerobic enzymes in this top 100 list differed between each species (see Supplementary Data 3). For *Skoliomonas litria*, *Skoliomonas* sp. GEMRC, and *Skoliomonas* sp. RCL this list included two or more copies of genes encoding PFL. For *Skoliomonas litria* and *Skoliomonas* sp. RCL, this list also included one copy each of a gene encoding PFO. *Skoliomonas litria*, in addition to PFL, had a copy of an ACS and a malic enzyme gene in its top 100 most highly-expressed gene set. *Skoliomonas* sp. GEMRC was the only organism with a SmsCB gene represented its top 100 expressed list; other BaSk members have comparatively lower expression levels of this gene.

### Implications for the evolution of MROs within the Metamonada supergroup

Based on our reconstructions of MRO properties on the metamonad tree, we suggest that the last common ancestor of the BaSk clade possessed an already highly reduced MRO compared to the common ancestor of fornicates + BaSk (Fig. 4). The barthelonids PCE and PAPO20 seem to have functionally reduced MROs, but their predicted MRO proteomes retain similar functions to other free-living metamonads such as the GCS and serine metabolism, hydrogen production and Fe/S cluster biogenesis by a simple ISC system. On the other hand, the MROs of *Skoliomonas* sp. RCL and *Skoliomonas* sp. GEMRC appear to be unique amongst metamonads, as we predict that the organelles only function in glycine and serine metabolism and are unlikely to produce H<sub>2</sub>. For *Skoliomonas litria*, the only proteins typically associated with mitochondria and MROs detected were a candidate MCP and a homolog of Sam50; of these, only the latter is exclusively associated with mitochondrial organelles in eukaryotes. The complete lack of all other proteins homologous to common MRO-localizing proteins indicates that an MRO may not even exist in this organism. In any case, this extreme reductive evolutionary path may relate to the acquisition of a simple fused SmsCB gene that was acquired by LGT in a common ancestor of BaSk. This acquisition then led to the simplification of the ISC system in the *Barthelona* clade, and the outright loss of the ISC system in the skoliomonads, as Fe/S proteins for organellar functions were no longer needed.

Regardless of the evolutionary forces involved, the BaSks demonstrate reduction in MRO functions that is independent of the losses seen in other metamonad groups, such as diplomonads and oxymonads<sup>1,6</sup>. This is remarkable given that many metamonads that display similar losses of MRO functions are parasitic or commensalistic and their nutrient-rich endobiotic lifestyle has often been assumed to be part of the explanation for their extreme MRO

proteome streamlining<sup>3,6</sup>. Instead, in the case of the free-living BaSk members, it appears that gain of a simple SMS system was amongst the first triggers leading to the extreme reductive trajectory of MRO evolution in this group.

## Discussion

We have generated high quality draft genomes, transcriptomes, and curated predicted proteome sets for a novel group of free-living metamonads. The placement of BaSks within the tree of Metamonada as the sister group to the Fornicata make them a key group to study in terms of genome evolution, LGT events, and adaptation to various environmental niches. These data place us several steps forward towards the goal of representing diverse and historically undersampled microbial eukaryotes in the tree of eukaryotic life.

Our findings demonstrate the diversity of MRO configuration and function amongst the metamonads. As a group, BaSks display a level of MRO reduction not typically seen in free-living metamonads and which is independent of the reduction found in the parasitic diplomonads. BaSk members share many interesting parallels with the MRO reduction and loss seen in other metamonads, as all BaSks possess a SMS Fe/S cluster system that has, in skoliomonads, replaced the canonical ISC system. This is the second discovered instance in metamonads of laterally acquired genes encoding a SUF-like system replacing the ISC system highlighting the role of adaptive LGT in MRO evolution. Based on these findings, we speculate that the last common ancestor of the BaSks possessed a compact, streamlined nuclear genome and a highly reduced MRO, and was markedly different from the last common ancestor of the fornicates. Further work is needed to elucidate the exact subcellular localization of several candidate MRO proteins including enzymes in the anaerobic ATP synthesis pathway and to determine the breadth of functions carried out by the MROs in this newly discovered group of metamonads. For *Skoliomonas litria*, which lacks all known mitochondrial pathways, it will be especially important to investigate the localization of its anaerobic ATP synthesis pathway, its MCP, and the Sam50/BamA-like protein to determine whether this free-living flagellate has completely dispensed with the organelle.

## Methods

### Cell cultivation and microscopy

Established cultures of isolates *Skoliomonas litria*, *Skoliomonas* sp. RCL and *Skoliomonas* sp. GEMRC<sup>22</sup> were maintained in variants of CR media<sup>63</sup>; *Skoliomonas litria* in 50 mL of 40 ppt CR with 3% LB, *Skoliomonas* sp. GEMRC in 50 mL of 40 ppt CR with 1% LB and 3 sterile wheat grains, *Skoliomonas* sp. RCL in 50 mL of 25 ppt CR with 1% LB and 3 sterile wheat grains. All were inoculated into fresh media every 7–10 days based on culture density, with a second inoculation after 24 h. Cultures of isolate ‘*Barthelona* sp.’ PCE<sup>8</sup> were maintained in sterile filtered seawater (Northwest Arm, Halifax, Nova Scotia) supplemented with 3% LB, and kept at 16°C. Cell growth was monitored weekly, and subculturing was done every 3–4 weeks depending on cell density by transferring around 20% of the old culture into the new culture.

Cells were imaged with differential interference contrast optics on a Zeiss Axiovert 200M microscope fitted with a AxioCam M5 camera (Carl Zeiss AG).

### Nucleic acid extraction and sequencing

For *Skoliomonas litria*, *Skoliomonas* sp. GEMRC, *Skoliomonas* sp. RCL, and *Barthelona* sp. PCE: cell material for DNA and RNA extraction was harvested from 1–2 L of dense cell culture. The cells were collected by centrifugation in 50 mL tubes for 8 min at 2000 × g and at 4°C. The pellets were then combined into two 15 mL tubes that were spun again as before. The recovered pellets were resuspended in sterile filtered (0.2 µm) spent culture media and layered over Histopaque-1077, and then subjected to centrifugation at 2000 × g for 20 min at room temperature. The protist-containing layer between the Histopaque

and the top media was collected, and the procedure was repeated for a total of two rounds of Histopaque separation. Finally, the cells were diluted by 5 mL (10 volumes) of sterile spent-media and pelleted for 8 min, at 2000 × g, 4°C. The cells were then resuspended in 2 mL of fresh media.

Genomic DNA was purified using a variety of techniques including traditional phenol-chloroform extraction, CTAB extraction, the QIA-GEN Genomic-tip kit, and the QIAGEN MagAttract HMW DNA kit. RNA was extracted using Trizol according to Invitrogen specifications. RNA and DNA were sent to Genome Quebec for library construction and NovaSeq 6000 Illumina sequencing. In the case of the RNA, poly-A selection was performed to enrich for eukaryotic reads. High molecular weight DNA from each species was used to construct a 1D ligation Nanopore library (LSK108 for *Barthelona* sp. PCE, LSK109 - EXP-NBD104 for *Skoliomonas litria* and *Skoliomonas* sp. GEMRC, and LSK308 for *Skoliomonas* sp. RCL) and was sequenced using a MinION flowcell (FLO-MIN106 (R9.4) for *Skoliomonas litria* and *Skoliomonas* sp. GEMRC, and FLO-MIN107 (R9.5) for *Skoliomonas* sp. RCL and *Barthelona* sp. PCE).

For *Barthelona* sp. PAP020, cells were cultured as described previously<sup>8</sup>. Cells were harvested by pelleting at 1500 rpm for 5 min at 4°C and were washed with PBS prior to phenol chloroform DNA extraction. The purified DNA was sequenced on an Illumina Hiseq 2500 PE using the Trueseq library prep.

### Genome/transcriptome assembly and gene annotation

For *Skoliomonas litria*, *Skoliomonas* sp. GEMRC, *Skoliomonas* sp. RCL, and *Barthelona* sp. PCE, Illumina NovaSeq reads were trimmed using Trimmomatic v0.36<sup>64</sup>. RNASeq reads were assembled using Trinity v2.6.6<sup>65</sup> and checked for multiplexing contamination using WinstonCleaner (<https://github.com/kolecko007/WinstonCleaner>). Bacterial contamination was removed from the Trinity assemblies using Anvi'o v5<sup>66</sup>.

Raw MinION sequencing data was basecalled using Albacore v2.1.3 (Albacore has since been replaced by Guppy—see ref. <sup>67</sup>). These data were assembled using Flye v2.3<sup>68</sup>, Raven v0<sup>69</sup> Canu v1.7<sup>70</sup> and ABrujin v2.2b<sup>71</sup>. The resulting assemblies were evaluated for completeness and contiguity by comparing assembly size, contig number, N50 and L50 values, and ALE scores<sup>72</sup>. The ABrujin assembly was chosen moving forward for *Skoliomonas litria*, *Skoliomonas* sp. RCL and *Barthelona* sp. PCE. For *Skoliomonas* sp. GEMRC, the Canu assembly was used. Raw MinION output was used to improve the base-call accuracy of these assemblies using Nanopolish version 0.8.4 (<https://github.com/jts/nanopolish>) and Illumina reads were used to error-correct the assemblies using Unicycler version 0.4.3<sup>73</sup>. Bacterial contamination was assessed and removed from the resulting contigs using read coverage, GC content, and BLAST search results. The Eukfinder workflow was used to remove prokaryotic contaminants from the BaSk genome assemblies<sup>23</sup>.

Gene prediction utilized a pipeline described in ref. <sup>25</sup>. Briefly, RepeatMasker<sup>74</sup> was first used to mask the repetitive regions of the genome. RNA-seq reads were then mapped to the genome using HISAT2<sup>75</sup>. GeneMark-ET<sup>76</sup> then utilized this mapping information to generate gene coordinates and intron splice sites, creating a training set of genes. This training set was then used to train AUGUSTUS<sup>77</sup>, which generated the gene predictions. PASA<sup>78</sup> was then used to validate the coordinates of these gene models. The predicted protein set was checked for completeness using the obd9 BUSCO v3.0.1 database<sup>79</sup>.

To estimate genome completeness we used Merqury v.1.0<sup>24</sup> with a kmer size of 17 (estimated using Merqury). The Illumina reads used for the Merqury analysis were decontaminated using the Eukfinder workflow<sup>23</sup> before genome completeness estimation. Additionally, RNA-seq mapping was used as a measure of genome completeness, and RNA-seq reads were decontaminated using the same method mentioned used for the DNA Illumina reads.

To construct the *Barthelona* sp. PAP020 genome, Illumina reads were assembled using Spades<sup>80</sup> with default parameters. Bacterial reads were identified by metaBAT<sup>81</sup> and CheckM<sup>82</sup>. Once the identified bacterial reads were removed, the remaining Illumina reads were then used to assemble a new genome. This process was repeated until no bacterial contamination was detected. The final contigs were verified by checking kmer content and were also checked against the *Barthelona* sp. PAP020 transcriptome<sup>8</sup>.

### Gene expression analysis

Expression of predicted protein-coding genes was quantified by mapping RNASeq reads to their respective long-read genome using HISAT2<sup>75</sup>. Using this mapping information and gene model coordinates, the expression level was normalized using Cufflinks (Cuffquant and Cuffnorm with default settings)<sup>83</sup> to compute a FPKM value for each gene.

### Phylogenomic analysis

The phylogenomic dataset was constructed based on a previously published set of 351 highly conserved protein orthologs<sup>34</sup>, with 177 orthologs excluded due to potential deep paralogies or insufficient coverage across pertinent species in the analysis. Each of the 174 remaining ortholog alignments were used to construct single-protein phylogenies, and were used to screen out non-orthologous data. The final 174 alignments were realigned using mafft v7.310 einsi, trimmed using BMGE v1.0 (default settings)<sup>84</sup>, and were concatenated together and used to infer a maximum-likelihood tree under the LG + C60 + F+Γ model of evolution with IQTree v1.5.5<sup>85</sup>. A PMSF model based on this tree and mixture model<sup>86</sup> was estimated and used to calculate branch supports using non-parametric bootstrapping, approximate Bayes (aBayes<sup>87</sup>); and approximate likelihood-ratio support (aLRT<sup>88</sup>).

### Inventory of MRO and anaerobic metabolism proteins

Proteins of interest were searched for in the predicted proteome, transcriptome, and genome of each species using both BLAST (using blastp, tblastn and blastx) and HMM profile searching using Hmmer v3.1b2 (hmmer.org). For each search, queries included both orthologous sequences from closely related metamonads and, once identified by these methods, the orthologs from other BaSks. The protein sets of the BaSk taxa were also used as queries in BLAST searches against the mitochondrial proteomes from *Andalucia godoyi* (Genbank accession VRVR00000000<sup>89</sup>), *Tetrahymena thermophila* (predicted protein Genbank accession EAR80512 to EAS07932<sup>90,91</sup>), *Arabidopsis thaliana* (NCBI Bioproject PRJNA10719 and SAMN03081427<sup>92,93</sup>), *Homo sapiens*<sup>94</sup> (mitochondrial proteins from UniProt (GO:0005739)), *Acanthamoeba castellanii* (NCBI Bioproject PRJNA599339 and PRJNA487265<sup>95,96</sup>), and *Saccharomyces cerevisiae* (<https://www.yeastgenome.org/><sup>97</sup>). The top five hits for each query sequence were examined manually. Proteins relating to mitochondrial protein translocation were searched for using both general HMM profiles<sup>1</sup> and profiles tailored for searching in metamonads. The latter profiles were created by retrieving the profiles from Pfam for each protein of interest and then retrieving the best hits using the Pfam profile against the several publicly available metamonad sequencing data databases with hmmsearch. Using an evalue cutoff of < 0.01, metamonad sequences were added to the corresponding Pfam seed alignment using the mafft -add function. If there were more than two detected sequences from metamonads, an alignment was made with these sequences and was then used to build a profile to search for more hits. Each candidate protein identified was used to reciprocally search the BLAST database to confirm protein identity. In ambiguous cases, the Pfam and InterPro databases were searched to analyze protein domain structure.

Phylogenies for key proteins identified were constructed to aid in identification in cases with deep paralogy (Hsp70 and Cpn60). Phylogenies were also made to identify the origin of enzymes that act in

anaerobic metabolism. Alignments<sup>1,98</sup> were utilized as a starting point, and we added the corresponding BaSk sequences. Where relevant, the protein sequences of the top 10 BLAST hits from each BaSk homolog were also added to the appropriate alignments. In cases where no starting alignments were available, a preliminary alignment was created by retrieving the top 1000 blast hits and reducing the sequences to 100-120 sequences using CD-HIT v.4.6 with recommended word sizes. These sequences were aligned using mafft einsi with default settings and were trimmed using trimal v1.4.rev15<sup>99</sup> with the -gappyout setting. These phylogenies were inferred using the LG + C20 + F+Γ model of evolution with IQTree.

### Mitochondrion-related organelle protein localization predictions

Putative mitochondrial targeted proteins and mitochondrial targeting sequences (MTS) were predicted using the following mitochondrial targeting peptide prediction programs: TargetP v1.1b<sup>100</sup>, MitoProt II v1.101<sup>101</sup>, and MitoFates v1.1<sup>102</sup>. Additionally, the predicted proteins were used to search both the nr database and a custom database of known mitochondrial proteins.

### Protein structure prediction and alignment

The structure of candidate Sam50/BamA-like proteins in *Skoliomonas litria* and *Skoliomonas* sp. GEMRC were predicted using the ColabFold v1.5.2 server<sup>103</sup>. The β-barrel portion of the *Skoliomonas litria* Sam50/BamA-like protein was aligned to the yeast Sam50 protein structure<sup>54</sup> using the PyMOL “cealign” function.

### SMSB (SUFCB) phylogeny

The SmsCB fusion protein identified in each BaSk member was added to previously constructed alignments<sup>13</sup> of SmsC and SmsB/D. This dataset was aligned using mafft einsi with default settings and was trimmed by hand. The dataset was then concatenated together, and the phylogeny was inferred using the same methods as the single protein phylogenies mentioned above.

### M16 protease phylogeny

The MPP candidate proteins identified in the skoliomonads were added to the previously constructed alignments of M16 proteases<sup>51</sup>, which include SPP, PreP, MPP-alpha and MPP-beta sequences. In addition, we used the protein sequences of the two MPP candidates of *Skoliomonas litria* to BLAST search nr, and added the top 20 hits for each candidate to the alignments. This dataset was aligned using mafft einsi with default settings and was trimmed by trimal as above. After trimming, we removed the duplicated protease domains in four domain proteins, relative to their position when aligned to two domain proteins. This alignment was then inferred using the same methods as the single protein phylogenies described above. The domain structures displayed to the right of the phylogeny in Supplementary Figure 7 represents the predicted protein domains of the skoliomonads (in red) and other major clades within the tree (in grey) as predicted by the InterProScan online tool (see <https://www.ebi.ac.uk/interpro/>), which utilizes the InterPro 98.0 database<sup>104</sup>.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

Genomes and gene annotations (where applicable) were deposited in the NIH GenBank database and can be found under BioProject ID PRJNA949547. Genomes, gene annotations, alignments, HMM profiles, and other datasets used in this study are available to download from Dryad [<https://doi.org/10.5061/dryad.73n5tb331>].

## References

1. Leger, M. M. et al. Organelles that illuminate the origins of *Trichomonashydrogenosomes* and *Giardiamitosomes*. *Nat. Ecol. Evol.* **1**, 1–15 (2017).
2. Schneider, R. E. et al. The *Trichomonas vaginalis* hydrogenosome proteome is highly reduced relative to mitochondria, yet complex compared with mitosomes. *Int. J. Parasitol.* **41**, 1421–1434 (2011).
3. Stairs, C. W., Leger, M. M. & Roger, A. J. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140326 (2015).
4. Jedelský, P. L. et al. The minimal proteome in the reduced mitochondrion of the parasitic protist *Giardia intestinalis*. *PLoS One* **6**, 15–21 (2011).
5. Sánchez, L. B., Galperin, M. Y. & Müller, M. Acetyl-CoA synthetase from the amitochondriate eukaryote *Giardia lamblia* belongs to the newly recognized superfamily of acyl-CoA synthetases (nucleoside diphosphate-forming). *J. Biol. Chem.* **275**, 5794–5803 (2000).
6. Karnkowska, A. et al. A eukaryote without a mitochondrial organelle. *Curr. Biol.* **26**, 1274–1284 (2016).
7. Müller, M. et al. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol. Mol. Biol. Rev.* **76**, 444–495 (2012).
8. Yazaki, E. et al. Barthelomids represent a deep-branching Metamonad clade with mitochondrion-related organelles generating no ATP. *Proc. R. Soc.* <https://doi.org/10.1101/805762> (2020).
9. Stairs, C. W. et al. Anaeramoebae are a divergent lineage of eukaryotes that shed light on the transition from anaerobic mitochondria to hydrogenosomes. *Curr. Biol.* **31**, 5605–5612.e5 (2021).
10. Zitek, J. et al. Reduced mitochondria provide an essential function for the cytosolic methionine cycle. *Curr. Biol.* **32**, 5057–5068 (2022).
11. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification of mitochondria. *Curr. Biol.* **32**, R1177–R1192 (2017).
12. Stairs, C. W., Roger, A. J. & Hampl, V. Eukaryotic pyruvate formate lyase and its activating enzyme were acquired laterally from a firmicute. *Mol. Biol. Evol.* **28**, 2087–2099 (2011).
13. Leger, M. M., Eme, L., Hug, L. A. & Roger, A. J. Novel hydrogenosomes in the microaerophilic jakobid *Stygiella incarcerata*. *Mol. Biol. Evol.* **33**, 2318–2336 (2016).
14. Stairs, C. W. et al. Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. *Elife* **7**, 1–23 (2018).
15. Jiménez-González, A., Xu, F. & Andersson, J. O. Lateral acquisitions repeatedly remodel the oxygen detoxification pathway in diplomonads and relatives. *Genome Biol. Evol.* **11**, 2542–2556 (2019).
16. Gawryluk, R. M. R. & Stairs, C. W. Diversity of electron transport chains in anaerobic protists. *Biochim. Biophys. Acta Bioenerg.* **1862**, 148334 (2021).
17. Garcia, P. S. et al. An early origin of iron–sulfur cluster biosynthesis machineries before earth oxygenation. *Nat. Ecol. Evol.* **6**, 1564–1572 (2022).
18. Nýltová, E. et al. Lateral gene transfer and gene duplication played a key role in the evolution of *Mastigamoeba balamuthi* hydrogenosomes. *Mol. Biol. Evol.* **32**, 1039–1055 (2015).
19. Záhonová, K. et al. Anaerobic derivates of mitochondria and peroxisomes in the free-living amoeba *Pelomyxa schiedti* revealed by single-cell genomics. *BMC Biol.* **20**, 1–16 (2022).
20. Stairs, C. W. et al. A SUF Fe-S cluster biogenesis system in the mitochondrion-related organelles of the anaerobic protist *Pygsuia*. *Curr. Biol.* **24**, 1176–1186 (2014).
21. Vacek, V. et al. Fe-S cluster assembly in oxymonads and related protists. *Mol. Biol. Evol.* **35**, 2712–2718 (2018).
22. Eglit, Y., Williams, S. K., Roger, A. J. & Simpson, A. G. B. Characterization of *Skoliomonas* gen. nov., a haloalkaliphilic anaerobe related to barthelomids (Metamonada). *J. Eukaryot. Microbiol.* <https://doi.org/10.1111/jeu.13048> (2024).
23. Zhao, D., Salas-leiva, D. E., Williams, S. K. & Dunn, K. A. Eukfinder: a pipeline to retrieve microbial eukaryote genomes from metagenomic sequencing data. *bioRxiv* <https://doi.org/10.1101/2023.12.28.573569> (2023).
24. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 1–27 (2020).
25. Salas-Leiva, D. E. et al. Genomic analysis finds no evidence of canonical eukaryotic DNA processing complexes in a free-living protist. *Nat. Commun.* **12**, 6003 (2021).
26. Schömknecht, G. et al. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**, 1207–1210 (2013).
27. Goffeau, A. A. et al. Life with 6000 genes. *Science* **274**, 546, 563–7 (1996).
28. Derelle, E. et al. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl Acad. Sci. USA* **103**, 11435–11436 (2006).
29. Xu, F., Jex, A. & Svärd, S. G. A chromosome-scale reference genome for *Giardia intestinalis* WB. *Sci. Data* **7**, 1–8 (2020).
30. Manzano-Marín, A. & Latorre, A. Snapshots of a shrinking partner: genome reduction in *Serratia symbiotica*. *Sci. Rep.* **6**, 1–11 (2016).
31. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).
32. Fulnečková, J. et al. A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. *Genome Biol. Evol.* **5**, 468–483 (2013).
33. Kang, S. et al. Between a pod and a hard test: the deep evolution of Amoebae. *Mol. Biol. Evol.* **34**, 2258–2270 (2017).
34. Brown, M. W. et al. Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biol. Evol.* **10**, 427–433 (2018).
35. Simpson, A. G. B. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *Int. J. Syst. Evol. Microbiol.* **53**, 1759–1777 (2003).
36. Kikuchi, G. The glycine cleavage system: composition, reaction mechanism, and physiological significance. *Mol. Cell. Biochem.* **21**, 1154–1157 (1973).
37. Zubáčová, Z. et al. The mitochondrion-like organelle of *Trimastix pyriformis* contains the complete glycine cleavage system. *PLoS One* **8**, 1–9 (2013).
38. Jerlström-Hultqvist, J. et al. Hydrogenosomes in the diplomonad *Spiرونucleus salmonicida*. *Nat. Commun.* **4**, 1–9 (2013).
39. Mukherjee, M., Brown, M. T., McArthur, A. G. & Johnson, P. J. Proteins of the glycine decarboxylase complex in the hydrogenosome of *Trichomonas vaginalis*. *Eukaryot. Cell* **5**, 2062–2071 (2006).
40. Nýltová, E., Smutná, T., Tachezy, J. & Hrdý, I. OsmC and incomplete glycine decarboxylase complex mediate reductive detoxification of peroxides in hydrogenosomes of *Trichomonas vaginalis*. *Mol. Biochem. Parasitol.* **206**, 29–38 (2016).
41. Motyčková, A. et al. Adaptation of the late ISC pathway in the anaerobic mitochondrial organelles of *Giardia intestinalis*. *PLoS Pathog.* **19**, 1–24 (2023).
42. Lill, R. & Freibert, S. A. Mechanisms of mitochondrial iron-sulfur protein biogenesis. *Annu. Rev. Biochem.* **89**, 471–499 (2020).
43. Ohnishi, T. Iron-sulfur clusters/semiquinones in complex I. *Biochim. Biophys. Acta Bioenerg.* **1364**, 186–206 (1998).
44. Pyrih, J. et al. Minimal cytosolic iron-sulfur cluster assembly machinery of *Giardia intestinalis* is partially associated with mitochondria. *Mol. Microbiol.* **102**, 701–714 (2016).
45. Žářský, V. et al. The *Mastigamoeba balamuthi* genome and the nature of the free-living ancestor of entamoeba. *Mol. Biol. Evol.* **38**, 2240–2259 (2021).

46. Anwar, S. et al. Interaction between Nbp35 and Cfd1 proteins of cytosolic Fe-S cluster assembly reveals a stable complex formation in *Entamoeba histolytica*. *PLoS One* **9**, 12–14 (2014).
47. Yubuki, N. et al. Ancient adaptive lateral gene transfers in the symbiotic *Opalina-Blastocystis* stramenopile lineage. *Mol. Biol. Evol.* **37**, 651–659 (2020).
48. Tsiaousis, A. D. et al. Evolution of Fe/S cluster biogenesis in the anaerobic parasite *Blastocystis*. *Proc. Natl Acad. Sci. USA* **109**, 10426–10431 (2012).
49. Liu, Y., Sieprawska-Lupa, M., Whitman, W. B. & White, R. H. Cysteine is not the sulfur source for iron-sulfur cluster and methionine biosynthesis in the methanogenic archaeon *Methanococcus maripaludis*. *J. Biol. Chem.* **285**, 31923–31929 (2010).
50. Nelson, R. J., Ziegelhoffer, T., Nicolet, C., Werner-Washburne, M. & Craig, E. A. The translation machinery and 70 kd heat shock protein cooperate in protein synthesis. *Cell* **71**, 97–105 (1992).
51. Garrido, C., Wollman, F. A. & Lafontaine, I. The evolutionary history of peptidases involved in the processing of organelle-targeting peptides. *Genome Biol. Evol.* **14**, 1–18 (2022).
52. Mazurek, M. P. et al. Molecular origin of plasma membrane citrate transporter in human prostate epithelial cells. *EMBO Rep.* **11**, 431–437 (2010).
53. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
54. Takeda, H. et al. Mitochondrial sorting and assembly machinery operates by β-barrel switching. *Nature* **590**, 163–169 (2021).
55. Pyrihová, E. et al. A single tim translocase in the mitosomes of *Giardia intestinalis* illustrates convergence of protein import machines in anaerobic eukaryotes. *Genome Biol. Evol.* **10**, 2813–2822 (2018).
56. Tanifuiji, G. et al. The draft genome of *Kipferlia bialata* reveals reductive genome evolution in fornicate parasites. *PLoS One* **13**, 1–18 (2018).
57. Garg, S. et al. Conservation of transit peptide-Independent protein import into the mitochondrial and hydrogenosomal matrix. *Genome Biol. Evol.* **7**, 2716–2726 (2015).
58. Mentel, M., Zimorski, V., Haferkamp, P., Martin, W. & Henze, K. Protein import into hydrogenosomes of *Trichomonas vaginalis* involves both N-terminal and internal targeting signals: a case study of thioredoxin reductases. *Eukaryot. Cell* **7**, 1750–1757 (2008).
59. Sanchez, L. B. & Müller, M. Purification and characterization of the acetate forming enzyme, acetyl-CoA synthetase (ADP-forming) from the amitochondriate protist, *Giardia lamblia*. *FEBS Lett.* **378**, 240–244 (1996).
60. Van Grinsven, K. W. A. et al. Acetate:succinate CoA-transferase in the hydrogenosomes of *Trichomonas vaginalis*: identification and characterization. *J. Biol. Chem.* **283**, 14111–14118 (2008).
61. Lahti, C. J., D’Oliveira, C. E. & Johnson, P. J. β-Succinyl-coenzyme a synthetase from *Trichomonas vaginalis* is a soluble hydrogenosomal protein with an amino-terminal sequence that resembles mitochondrial presequences. *J. Bacteriol.* **174**, 6822–6830 (1992).
62. Dyall, S. D. et al. Non-mitochondrial complex I proteins in a hydrogenosomal oxidoreductase complex. *Nature* **431**, 1103–1107 (2004).
63. Gigeroff, A. S., Eglit, Y. & Simpson, A. G. B. Characterisation and cultivation of new lineages of colponemids, a critical assemblage for inferring alveolate evolution. *Protist* **174**, 125949 (2023).
64. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
65. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with trinity. *Nat. Protoc.* **8**, 1494–1512 (2013).
66. Eren, A. M. et al. Community-led, integrated, reproducible multi-omics with anvi’o. *Nat. Microbiol.* **6**, 3–6 (2021).
67. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 1–10 (2019).
68. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
69. Vaser, R. & Šikić, M. Time- and memory-efficient genome assembly with Raven. *Nat. Comput. Sci.* **1**, 332–336 (2021).
70. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv* **25**, 1–2 (2016).
71. Lin, Y. et al. Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl Acad. Sci. USA* **113**, E8396–E8405 (2016).
72. Clark, S. C., Egan, R., Frazier, P. I. & Wang, Z. ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* **29**, 435–443 (2013).
73. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, 1–22 (2017).
74. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
75. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
76. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42**, 1–8 (2014).
77. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, 309–312 (2004).
78. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
79. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. Genome analysis BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
80. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes de novo assembler. *Curr. Protoc. Bioinforma.* **70**, 1–29 (2020).
81. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **2015**, 1–15 (2015).
82. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
83. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
84. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
85. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
86. Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67**, 216–235 (2018).
87. Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy,

- power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* **60**, 685–699 (2011).
88. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
89. Gray, M. W. et al. The draft nuclear genome sequence and predicted mitochondrial proteome of *Andalucia godoyi*, a protist with the most gene-rich and bacteria-like mitochondrial genome. *Curr. Biol.* **3**, 1–35 (2019).
90. Eisen, J. A. et al. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* **4**, 1620–1642 (2006).
91. Smith, D. G. S. et al. Exploring the mitochondrial proteome of the ciliate protozoon *Tetrahymena thermophila*: direct analysis by tandem mass spectrometry. *J. Mol. Biol.* **374**, 837–863 (2007).
92. Cheng, C. Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804 (2017).
93. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
94. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **412**, 565–566 (2001).
95. Gawryluk, R. M. R., Chisholm, K. A., Pinto, D. M. & Gray, M. W. Compositional complexity of the mitochondrial proteome of a unicellular eukaryote (*Acanthamoeba castellanii*, supergroup Amoebozoa) rivals that of animals, fungi, and plants. *J. Proteom.* **109**, 400–416 (2014).
96. Matthey-Doret, C. et al. Chromosome-scale assemblies of *Acanthamoeba castellanii* genomes provide insights into *Legionella pneumophila* infection-related chromatin reorganization. *Genome Res.* **32**, 1698–1710 (2022).
97. Engel, S. R. et al. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 Gene. Genome. Genet.* **4**, 389–398 (2014).
98. Tsaousis, A. D., Gentekaki, E., Eme, L., Gaston, D. & Roger, A. J. Evolution of the cytosolic iron-sulfur cluster assembly machinery in *Blastocystis* species and other microbial eukaryotes. *Eukaryot. Cell* **13**, 143–153 (2014).
99. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
100. Armenteros, J. J. A. et al. Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance* **2**, 1–14 (2019).
101. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241**, 779–786 (1996).
102. Fukasawa, Y. et al. MitoFates: Improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol. Cell. Proteom.* **14**, 1113–1126 (2015).
103. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
104. Paysan-Lafosse, T. et al. InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).

## Acknowledgements

The majority of work was supported by Foundation grant (FRN-142349) from the Canadian Institutes for Health Research (CIHR) awarded to

Andrew J. Roger. Shelby K. Williams was supported by a graduate scholarship from the Killam Foundation.

## Author contributions

Y.E. provided cultures and conducted light microscopy analyses. S.K.W., J.J.H. and R.J.O. conducted the experimental portions of the genomics and transcriptomics. S.K.W., J.J.H., R.J.O. completed the genome and transcriptome assemblies and decontamination. D.E.S.-L. and B.C. conducted computational gene prediction and S.K.W. manually refined the predictions. All remaining bioinformatic analyses of the genome/transcriptome data and phylogenetic analyses were conducted by S.K.W. in consultation with A.J.R., J.J.H. and C.W.S. except for analyses of the M16 family proteases that were conducted by T.N.A. and N.M. in consultation with S.K.W. S.K.W. drafted the manuscript and figures in consultation with A.J.R. All coauthors assisted in editing the manuscript. A.J.R. and A.G.B.S. initially conceived the study.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-50991-w>.

**Correspondence** and requests for materials should be addressed to Andrew J. Roger.

**Peer review information** *Nature Communications* thanks Julius Lukes, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024